# Colonel Blotto in the Phishing War

Pern Hui Chia[1] and John Chuang[2]

[1] Centre for Quantifiable Quality of Service (Q2S)*, NTNU
[2] School of Information, UC Berkeley
chia@q2s.ntnu.no, chuang@ischool.berkeley.edu

**Abstract.** Phishing exhibits characteristics of asymmetric conflict and guerrilla warfare. Phishing sites, upon detection, are subject to removal by takedown specialists. In response, phishers create large numbers of new phishing attacks to evade detection and stretch the resources of the defenders. We propose the Colonel Blotto Phishing (CBP) game, a two-stage Colonel Blotto game with endogenous dimensionality and detection probability. We find that the optimal number of new phishes to create, from the attacker's perspective, is influenced by the degree of resource asymmetry, the cost of new phishes, and the probability of detection. Counter-intuitively, we find that it is the less resourceful attacker who would create more phishing attacks in equilibrium. And depending on the detection probability, an attacker will vary his strategies to either create even more phishes, or to focus on raising his resources to increase the chance he will extend the lifetime of his phishes. We discuss the implications to anti-phishing strategies and point out that the game is also applicable to web security problems more generally.

**Keywords:** Phishing, Economics, Colonel Blotto, Web Security

## 1 Introduction

Phishing, among other web security issues, has remained a tricky problem today. While it is non-trivial to measure the exact financial losses due to phishing, and that many estimated loss figures appear overstated [9], the damage inflicted by phishing activities is never negligible. Realizing that technical sophistication alone will not be sufficient to fend off phishing activities, over the past few years, researchers have started to look at the ecosystem and modi operandi of phishing activities.

McGrath and Gupta found that phishers misuse free web hosting services and URL-aliasing services, and that phishing domains are hosted across multiple countries with a significant percentage of hosts belonging to residential customers [13]. Moore and Clayton identified different types of phishing attacks

* Centre for Quantifiable Quality of Service in Communication Systems (Q2S), Centre of Excellence, appointed by the Research Council of Norway, is funded by the Research Council, Norwegian Uni. of Science and Technology (NTNU) and UNINETT. http://www.q2s.ntnu.no

according to the way a phishing site is hosted [16]. The most common hosting vectors were found to be compromised web servers and free web-hosting services. While system admins and hosting companies are usually cooperative and quick to take down the phishing pages once notified, noticing them in the first place is challenging [16]. Moreover, victim servers were found to be re-compromised by the attackers to host phishing pages as the vulnerabilities of the servers remain unpatched [17]. Two notorious gangs, known as 'Rock Phish' and 'Avalanche'[3] even showed much technical sophistication in their massive and concerted phishing attacks. Both gangs exploited malware-infested machines and the fast flux method (mapping the domain name to different IP addresses (of different bots) by changing the DNS records in a high frequency) to extend the lifetime of a phishing site. Taking down the phishing pages from a large number of bots is extremely difficult, especially when the ISPs have only limited control and responsibility over malware-infested machines. This forces the defender to take-down the phishing sites by suspending the phishing domain names with the help from registrars and registries.

The above highlights several important challenges in defending against phishing activities. First, it is challenging to detect all phishing attacks out there. Second, taking down phishing attacks that have been identified (e.g., to remove the phishing sites, or to ensure that a vulnerable web server is patched to prevent re-compromise) is also non-trivial. The situation is worsened by a lack of information sharing in the anti-phishing industry [16]. Meanwhile, despite a spike in the count of phishing attacks[4] in 2009 due to the Avalanche gang [2], the number of unique phishing domains found (per six months) has remained steady at around 30,000 over the past few years, except in the second half of 2010 where 43,000 unique phishing domain names were recorded partly due to new data inputs from the China Internet Network Information Center (CNNIC) who operates the `.cn` registry [3].[5] This suggests that the phishers do factor in the cost consideration when carrying out phishing attacks.

Different from prior studies that have largely taken the empirical approach, we propose in this work a theoretical model to aid researchers and policymakers in better analyzing the different aspects of phishing defense. We build on the Colonel Blotto game, an old but interesting game that has been largely neglected due to its complexity, until the recent work by Roberson [18] which gives a complete characterization to the unique equilibrium payoffs of a two-player asymmetric Colonel Blotto game. The game is particularly suitable to capture the resource allocation problem between a phisher and a defender with asymmetrical resources. In addition to mapping the phishing problem into the Colonel

[3] An account of the modi operandi of the Rock Phish and Avalanche gangs can be found in [14] and [2] respectively.

[4] An attack is defined by Anti-Phishing Working Group (APWG) as a unique phishing site targeting a specified brand.

[5] Measurement of unique phishing attacks, uptime of phishing sites and in-depth surveys on the trends and domain name use by phishing sites can be found in a series of reports (e.g., [2,3]) by the APWG on `http://www.antiphishing.org`.

Blotto game, our model extends the two-stage Colonel Blotto game in [10] to include a detection probability to factor in the consideration of asymmetric information that not all phishes will be known to the defender. We regard the defender in this work as a takedown company (e.g., MarkMonitor[6], BrandProtect[7] and Internet Identity[8]) that has been contracted by its clients (e.g., financial institutions, e-commerce services) to remove phishing sites that masquerade as the clients' legitimate sites. Although the defender is in a disadvantage position for not being able to detect all phishes that have been created, and that the attacker can always exploit the next weakest link whenever a phishing server is taken down, we expect that the defender can garner more resources than the attackers from the contract with its clients, plus the support from the ISPs, service providers, law enforcers, registrars and registries.

In the following, we first give a quick introduction to the Colonel Blotto game and related work in Section 2. We propose the Colonel Blotto Phishing (CBP) game in Section 3 to model phishing attack and defense. We present the results from our analysis based on the CBP model in Section 4. And lastly, we discuss the implications to the anti-phishing strategies in Section 5.

## 2 Background and Related Work

The *Colonel Blotto* game was first introduced in 1921 by Borel [6] as a two-player constant-sum game, where the players strategically distribute a fixed and *symmetrical* amount of resources over a finite number of $n$ contests (battlefields). The player who expends a higher amount of resources in a contest wins that particular battlefield, similar to an all-pay auction. The objective of the players is to maximize the number of battlefields won. Gross and Wagner [8] in 1950 described the game with *asymmetrical* resources between the two players, but have only solved the case where the number of battlefields $n = 2$.

The complexity for the case when there are $n \geq 3$ battlefields and the lack of pure strategies have arguably led to the Colonel Blotto game being largely neglected by the research community. A resurgence of interests in the Colonel Blotto game (e.g., [4,5,7,11,12,19]) follows the recent work by Roberson [18] which has successfully characterized the unique equilibrium payoffs for all configurations of resource asymmetry, and the equilibrium resource allocation strategies (for most configurations) of a constant-sum Colonel Blotto game with $n \geq 3$ battlefields. Roberson and Kvasov have later studied the non-constant-sum version in [19]. We summarize the main results from Roberson [18] below:

**Theorem 1** (case a, b and c correspond to Theorem 2, 3 and 5 in [18])
*Let $n$ denote the number of battlefields, while $R_w$ and $R_s$ denote the resources of the weak (w) and strong (s) players respectively such that $R_w \leq R_s$, the Nash*

[6] http://www.markmonitor.com
[7] http://www.brandprotect.com
[8] http://internetidentity.com

*equilibrium univariate distribution functions (for allocating resources to individual battlefields strategically), and the unique equilibrium payoffs (measured in the expected proportion on battlefields won), depending on the $\frac{R_w}{R_s}$ ratio and the number of battlefields $n$, are given in the following:*

*case a:* $\frac{2}{n} \leq \frac{R_w}{R_s} \leq 1$

*In the unique Nash equilibrium, player $w$ and $s$ allocate $x_j$ resources in each battlefield $j \in \{1, ..., n\}$ based on the following univariate distribution functions:*

$$F_{w,j}(x) = (1 - \tfrac{R_w}{R_s}) + \tfrac{nx}{2R_s}(\tfrac{R_w}{R_s}) \qquad , \ x \in [0, \tfrac{2R_s}{n}]$$
$$F_{s,j}(x) = \tfrac{nx}{2R_s} \qquad , \ x \in [0, \tfrac{2R_s}{n}]$$

*The unique equilibrium payoffs (expected proportions of battlefields won) of player $w$ and $s$ are independent of the number of battlefields, given as follows:*

$$\pi_w = \tfrac{R_w}{2R_s}$$
$$\pi_s = 1 - \tfrac{R_w}{2R_s}$$

*case b:* $\frac{1}{n-1} \leq \frac{R_w}{R_s} < \frac{2}{n}$

*In the unique Nash equilibrium, player $w$ and $s$ allocate $x_j$ resources in each battlefield $j \in \{1, ..., n\}$ based on the following univariate distribution functions:*

$$F_{w,j}(x) = (1 - \tfrac{2}{n}) + \tfrac{x}{R_w}(\tfrac{2}{n}) \qquad , \ x \in [0, R_w]$$
$$F_{s,j}(x) = \begin{cases} (1 - \frac{R_s}{nR_w})(\frac{2x}{R_w}) & , \ x \in [0, R_w) \\ 1 & , \ x \geq R_w \end{cases}$$

*The expected proportions of battlefields won by player $w$ and $s$ are as follows:*

$$\pi_w = \tfrac{2}{n} - \tfrac{2R_s}{n^2 R_w}$$
$$\pi_s = 1 - \tfrac{2}{n} + \tfrac{2R_s}{n^2 R_w}$$

*case c:* $\frac{1}{n} < \frac{R_w}{R_s} < \frac{1}{n-1}$

*In a Nash equilibrium, player $w$ allocates zero resources to $n-2$ of the battlefields, each randomly chosen with equal probability. On the remaining 2 battlefields, he randomizes the resource allocation over a set of bivariate mass points. On the other hand, player $s$ allocates $R_w$ resources to each of $n-2$ randomly chosen battlefields. On the remaining 2 battlefields, player $s$ also randomizes the resource allocation over a set of bivariate mass points. Let $m = \lceil \frac{R_w}{R_s - R_w(n-1)} \rceil$ such that $2 \leq m < \infty$, the unique expected proportions of battlefields won by player $w$ and $s$ are given as follows:*

$$\pi_w = \tfrac{2m-2}{mn^2}$$
$$\pi_s = 1 - \tfrac{2m-2}{mn^2}$$

Note that the univariate distribution functions constitute the players' mixed strategies in Nash equilibrium. The allocation of resources across the $n$ battlefields must additionally be contained in the set of all feasible allocations

$\left\{\mathbf{x} \in \mathbb{R}_+^n \middle| \sum_{j=1}^n x_{i,j} \leq R_i\right\}$ where $i = w, s$.[9] In general, player $s$ uses a stochastic 'complete coverage' strategy (which expends non-zero resources in all battlefields, and locks down in a random subset of battlefields by allocating $R_w$ resources to them in case b and c), while player $w$ uses a stochastic 'guerrilla warfare' strategy (which optimally abandons a random subset of the battlefields). Despite the resource asymmetry, player $w$ can expect to win a non-zero proportion of the battlefields, except in the case of $R_s \geq nR_w$, where the player $s$ can trivially lock down (win) all battlefields by allocating $R_w$ resources to each of them.

Note that also the proportion of battlefields won by the player $w$ is a function of $n$ in the case b and c of Theorem 1. In a recent work, Kovenock et al. [10] presented a two-stage Colonel Blotto game which endogenizes the dimensionality of the classic Colonel Blotto game, allowing the players to create additional battlefields in the additional 'pre-conflict' stage. They show that with such possibility, player $w$ will optimally increase the number of battlefields in the 'pre-conflict' stage, given a low battlefield creation cost, so to thin the defender's resources and reduce the number of battlefields player $s$ can lock down in the 'conflict' stage. We outline the main results from [10] below:

**Theorem 2** (see Theorem 2 in [10])
*In the pre-conflict stage of the game with $n_0$ initial battlefields and resource asymmetry that satisfies $\frac{1}{n_0-1} \leq \frac{R_w}{R_s} \leq 1$, assuming that the cost to create additional battlefields, c is strictly increasing and strictly convex, the optimal numbers of new battlefields that player w and s will create, $n_w^*$ and $n_s^*$ respectively, in the subgame perfect equilibrium, are given as follows:*

*<u>case a:</u> If $\frac{R_w}{R_s}$ satisfies $\frac{2}{n_0} \leq \frac{R_w}{R_s} \leq 1$, then $n_s^* = n_w^* = 0$.*

*<u>case b:</u> If $\frac{R_w}{R_s}$ satisfies $\frac{1}{n_0-1} \leq \frac{R_w}{R_s} < \frac{2}{n_0}$, then $n_s^* = 0$, and let $n_{wr} \in (0, \frac{2R_s}{R_w} - n_0)$ denotes the real number that solves:*

$$-\frac{2}{(n_0+n_{wr})^2} + \frac{4R_s}{R_w(n_0+n_{wr})^3} - c'_{n_{wr}} = 0$$

*then, $n_w^*$ is either $\lceil n_{wr} \rceil$ or $\lfloor n_{wr} \rfloor$ depending on which of it results in a higher utility for player w, given $n_s^* = 0$.*

Note that Theorem 2 has not formally treated the case $c$ of Theorem 1. The analysis of case $c$ will be more complicated as the expected proportion of battlefields won by both players have points of discontinuity, but the underlying intuition is the same as case $b$. [10] Note that also Theorem 2 assumes that the cost of creating additional battlefields is expended separately from the players' resources.

[9] We refer interested readers to Roberson [18] for proofs and details on how the equilibrium univariate distribution functions give a n-variate joint distribution function satisfying the constraint that $\sum_{j=1}^n x_{i,j} \leq R_i$ where $i = w, s$.

# 3 Modeling

With an introduction to the classic Colonel Blotto game and the extension with endogenous dimensionality, we are now ready to model the economics for phishing activities in this section. We will first apply the classic Colonel Blotto game to phishing attack and defense. Then, we will extend the game to model endogenous dimensionality following the two-stage Colonel Blotto game in [10], and asymmetric information using an additional detection probability to reflect that not all phishes will be known to the defender in practice.

## 3.1 Applying Colonel Blotto to Phishing

We map the basics the Colonel Blotto game in the context of phishing attack and defense in the following.

**Players.** Like the classic Colonel Blotto, we consider here a two-player constant-sum game between a phisher and a defender. We regard the defender here to be a takedown company such as MarkMonitor, BrandProtect and Internet Identity as aforementioned. The takedown company is contracted by its clients, including banks and popular brand owners, to remove phishing sites attacking the clients' brands. On the other hand, the phisher plays to keep alive the phishing sites, or to launch new attacks, to victimize as many users he can.

**Resources.** We assume the phisher to be the weak player ($w$) and the takedown company to be the strong player ($s$). Although this may be debatable, assuming such resource asymmetry is reasonable if we consider that takedown companies will usually maintain good contacts with and can thus get assistance from the ISPs, service providers, law enforcers, registrars and registries in the process of taking down the phishes. By resources, we thus mean not financial figures but mainly the *technologies*, *infrastructure* (e.g., access to a botnet), *time* and *manpower*.[10] Phisher's profitability is also not as lucrative as it appears in the news. A number of estimates on the losses due to phishing attacks have been criticized to be overstated [9]. The resources, $R_s$ and $R_w$ respectively, are finite with $R_s \geq R_w$. They are of the 'use-it-or-lose-it' nature, meaning that unused resources will give no value to the players in the end of the game.

**Battlefields.** We define a battlefield to be a unique phishing site (a fully qualified domain name or IP address, or a site on a shared hosting service) targeting a specific brand, following the definition of a phishing attack by APWG (see e.g., page 4 in [3]). Different URLs directing to the same phishing page, crafted to evade spam filters or to trick the URL-based anti-phishing toolbars, are considered the same battlefield. Defined this way, creating a battlefield hence involves some costs ranging from *low* (e.g., to register a subdomain on a shared hosting service, to copy the login page of a brand) to *high* (e.g., to register a new domain name, to compromise a vulnerable web server). In this paper, we use the terminologies 'a phish' and 'a phishing attack' interchangeably.

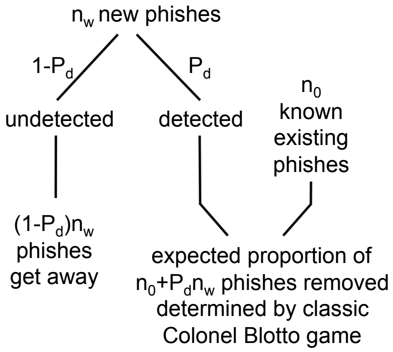
[10] Resource asymmetry should not be confused with asymmetry in coverage where the defender needs to protect all assets while the attacker can target any of them.

**Objectives & Contests.** We model the objective of the phisher and the defender to be maximizing the expected proportion of phishing attacks kept online and taken down, respectively. We consider that either the phisher or the defender can outperform the other party to win a battlefield by allocating more resources to it. And given that we have not factored in the uptime and the number of victims per attack in our model, we loosely define that a specific battlefield (phishing attack) is won by the phisher if the phish has a *long enough uptime*. For example, having the resources of a botnet infrastructure, an attacker can use 'fast-flux' IP addresses and malware-controlled proxies, to make it hard for the defender to take down the phishing server, prolonging the uptime of the phishes, as the defender will have to turn to the responsible registrar or registry to suspend the domain name. We elaborate on other tricks used by phishers, including the two infamous Rock Phish and Avalanche gangs, in Section 3.2.

Given the above configurations, we can already gain a number of useful insights. For example, we can expect that there will be always some phishes that will have long uptime unless that the defender is much more resourceful than the phisher (i.e., $R_s \geq nR_w$). However, the classic colonel blotto game alone does not describe the practical scenario quite yet. Why are there a large number of phishing attacks instead of just a few? Indeed, it is to the phisher's advantage to create an optimal number of additional phishes (battlefields), so to thin the defender's resources in removing each of them. Furthermore, how does the asymmetric information affect the strategies of the phisher? We extend the two-stage Colonel Blotto game in [10] to include an additional parameter, the expected probability of detection $P_d$, to reflect that not all phishes will be known to the defender – a major challenge in the anti-phishing industry. [16]

Table 1: The flow of the Colonel Blotto Phishing game.

| | Stage | Phisher ($w$) | Defender ($s$) |
|---|---|---|---|
| i) | create – detect | a. create and market $n_w^*$ new phishes<br>b. learn about detection | a. detect new phishes<br>b. publish findings |
| ii) | resist – takedown | c. expend $\varepsilon$ resources to undetected phishes, while allocating $R_w - \varepsilon$ resources to phishes known to the defender to resist removal | c. expend all $R_s$ resources strategically to remove the newly detected and known existing phishes in a promptly manner |



$n_w$ new phishes
1-$P_d$
$P_d$
undetected
detected
$n_0$ known existing phishes
(1-$P_d$)$n_w$ phishes get away
expected proportion of $n_0$+$P_d$$n_w$ phishes removed determined by classic Colonel Blotto game

Fig. 1: Expected proportion of phishes in different states.

### 3.2 The Colonel Blotto Phishing Game

We name our model as the Colonel Blotto Phishing (CBP) game. It consists of two stages: (i) create–detect, (ii) resist–takedown, similar to the 'pre-conflict' and 'conflict' stages in [10]. Table 1 summarizes the flow of the CBP game. We detail on the game stages in the following.

**Stage 1: Create–Detect.** We consider that game starts with the phisher having a number of phishes $n_0$ that are known to the defender, and both players are allowed to increase the dimensionality of the game by introducing new battlefields in the first stage. Obviously, the defender will not create any phishes. However, it is to the phisher's advantage to create a number of new phishing attacks $n_w$ so to stretch the defender's resources, in hope to increase the expected proportion of phishes that will stay online for more than a certain period of time. Hence, we have the total phishing attacks $n = n_0 + n_w$. We expect the phisher then advertises the newly created phishes through spams and online social networks.[11] We assume a linear cost $c$ for creating and advertising the new phishes; $c$ can be low or high depending on the way the phisher carries out the attack (e.g., through free subdomain services, paying for a newly registered domain, taking the effort to hack a vulnerable web server, and so on).

A new aspect we incorporate into the classic Colonel Blotto game is the situation where some of the newly created phishes might not be detected by the takedown company. We analyze both cases where the expected detection probability $P_d$ is (i) exogenously determined, and (ii) endogenously influenced by the number of new phishing attacks in Section 4. The expected proportions of phishes that trivially get away undetected, or that will possibly stay online long enough depending on the resource allocations of both the phisher and defender in the second stage, are depicted in Figure 1. In practice, takedown companies learn about new phishing attacks through their own infrastructures (e.g., spam

[11] McGrath and Gupta [13] observed that most domains created for phishing become active almost immediately upon registration.

filters) in addition to 'raw' feeds bought, negotiated or obtained from the ISPs or phishing clearinghouses, such as the APWG and PhishTank[12].

An assumption we make here is that the phisher will then learn about which of his phishes have been detected before proceeding to the next game stage. This is reasonable, regardless of whether the takedown company shares their detection results[13], as we expect that the phisher can achieve this using public clearinghouses (e.g., phishtank) or through anti-phishing APIs that come with modern browsers (e.g., Google Safe Browsing API[14] for FireFox and Chrome).

**Stage 2: Resist–Takedown.** Knowing the identity of the detected phishes $\mathbb{J}_d$, the optimal move for the phisher in the second stage is hence to expend all his resources strategically on the detected phishes only, so to resist the takedown process. Here, we assume that the resources (e.g., technologies, infrastructure, manpower) are of the 'use-it-or-lose-it' nature, typical to a constant-sum game. In other words, unused resources will give no value to the players. We further assume that the phisher will optimally allocate $\varepsilon \approx 0$ resources for the undetected phishes $j \notin \mathbb{J}_d$ given that the defender does not know about them. We note that this assumption is reasonable as the resources are finite.

We regard that either the phisher or the takedown company will 'succeed' with respect to a particular phishing attack depending on the amount of resources they put in: the player who expends more resources wins. Specifically, with $x_{i,j}$ and $x_{-i,j}$ denoting the amount of resources player $i \in \{w, s\}$ and his opponent puts into the phish attack $j$ respectively, the success of player $i$ at attack $j$ is given by:

$$\pi_{i,j}(x_{i,j}, x_{-i,j}) = \begin{cases} 0 & \text{if } x_{i,j} < x_{-i,j} \\ 1 & \text{if } x_{i,j} > x_{-i,j} \end{cases}$$

where in the case of $x_{w,j} = x_{s,j}$ (a tie), we assume that defender $s$ will succeed in taking down the attack promptly. As for undetected phishes, i.e., $\forall j \notin \mathbb{J}_d$, we regard that $x_{s,j} = 0$ and the phisher will trivially win the battlefield with $x_{w,j} = \varepsilon$ resources.

**Can the phisher still win in an already detected phish in practice?** While it may not be intuitive at first, the answer is 'yes' given our definition that a phishing attack is won by the phisher (defender) if the phish has an uptime more (less) than a certain threshold. The longer a phish can resist being removed, the more users could fall victim to it. While a weak phisher may simply abandon his phishes (given that he cannot win) when facing a much more resourceful defender (i.e., when $R_s \geq nR_w$), there have been practical examples of how a

[12] PhishTank – a community based phishing collator. `http://www.phishtank.com`

[13] Individual takedown companies often will validate the 'raw' URLs of potential phishes to remove false positives, and they might not voluntarily share their validated feeds for competitive advantages. Moore and Clayton showed how sharing of phishing data could have helped to halve the lifetime of phishes, translating to a potential loss mitigation of $330 million per year, based on data feeds from two takedown companies [15].

[14] `http://code.google.com/apis/safebrowsing/`

skilled phisher attempts to extend the lifetime of his phishes via different tricks. For example, a phisher may configure his phishes not to resolve on every access so to misguide the defender, but remain online to trick more users (see e.g., [3], footnote 5). The phisher may also temporarily remove the phishing pages from a compromised web server so to avoid further actions from the defender or admin (e.g., to patch up specific vulnerabilities) and re-plant the phishes at a later time. Indeed, APWG (see e.g., [3], footnote 5) finds that more than 10% of phishes are re-activated after being down for more than an hour. Moore and Clayton also found that 22% of all compromised web servers are re-compromised within 24 weeks to be used as the host for phishing sites [16].

With more resources, a phisher can even increase technical sophistication so to use malware-controlled proxies and fast-flux IP addresses as demonstrated the large-scale attacks by the infamous 'Rock Phish' and 'Avalanche' phishing gangs. The fast-changing nature of IP address that the phishing site resolving to indicate that the attacker has in control of a large number of compromised machines (bots) make it infeasible for the takedown company and the responsible ISPs to take the phishing servers offline promptly. Instead, the defender will have to work towards suspending the domain names in use, which could take a while if the responsible registrars are not responsive or have limited experience in abuse control. The 'Avalanche' gang was found to have exploited this; at the same time as they launched their massive attacks using domains bought from a few registrars (resellers), the gang scouted for other unresponsive registrars for future use (see page 7 of [2]). Meanwhile, in [14] Moore and Clayton found that the fast-flux phishing gang used 57 domain names and 4287 IP addresses for fast-flux phishing. The 1:75 skewed ratio is interesting as it suggests that the fast-flux phishing gang was highly resourceful (having access to a botnet infrastructure). However, we note that these resources are not unlimited. For example, the operations of the 'Avalanche' gang was disrupted as the security community affected a 'temporary' shut-down of the botnet infrastructure in Nov 2009 [2]. Later, although the gang managed to re-establish a new botnet, they were also found to prefer using their resources for a more profitable opportunity to distribute the Zeus malware, which has been designed to automate identity theft and facilitate unauthorized transactions. [3]

**Subgame Perfect Equilibrium.** We consider the objective of the phisher (the takedown company) is to maximize the proportion of phishes that he succeeds in keeping alive for a certain period (removing promptly), minus the cost for creating new phishing attacks. With $\mathbf{x}_i$ and $\mathbf{x}_{-i}$ denoting the resource allocations across all phishing sites by player $i \in \{w, s\}$ and his opponent respectively, the utility of player $i$ can be written as:

$$U_i(\{\mathbf{x}_i, n_i\}, \{\mathbf{x}_{-i}, n_{-i}\}) = \frac{1}{n}(\sum_{j \in \mathbb{J}_d} \pi_{i,j} + \sum_{j \notin \mathbb{J}_d} \pi_{i,j}) - cn_i$$

Note that $\mathbf{x}_i$ and $\mathbf{x}_{-i}$ must be contained in the set of all feasible allocations, given by $\{\mathbf{x}_i \in \mathbb{R}^n_+ | \sum_{j=1}^n x_{i,j} \leq R_i\}$.

The optimal number of new phishes to create $n_i^*$ and the optimal utility $U_i^*$ in subgame perfect equilibrium can be obtained by backward induction. First, we can work out the expected proportion of success of each player in the 'resist–takedown' stage based on Theorem 1 and the fact that a fraction of phishes will get away undetected as given by $P_d$. Then, returning to the 'create–detect' stage, the optimization problem of the phisher becomes:

$$\begin{aligned}\max_{n_w} E(U_w|n_w) &= \frac{1}{n}E(\sum_{j\in\mathbb{J}_d}\pi_{w,j}) + \frac{(1-P_d)n_w}{n} - cn_w \\ &= \frac{n_d}{n}E(\pi_w) + \frac{(1-P_d)n_w}{n} - cn_w\end{aligned}$$

where

$$\begin{aligned}E(\pi_w) &= \begin{cases}\frac{R_w}{2R_s} & \text{if } 1 \geq \frac{R_w}{R_s} \geq \frac{2}{n_d} \\ \frac{2}{n_d} - \frac{2R_s}{(n_d)^2R_w} & \text{if } \frac{2}{n_d} \geq \frac{R_w}{R_s} \geq \frac{1}{n_d-1} \\ 0 & \text{if } \frac{1}{n_d} \geq \frac{R_w}{R_s}\end{cases} \\ n_d &= P_d n_w + n_0 \\ n &= n_w + n_0\end{aligned}$$

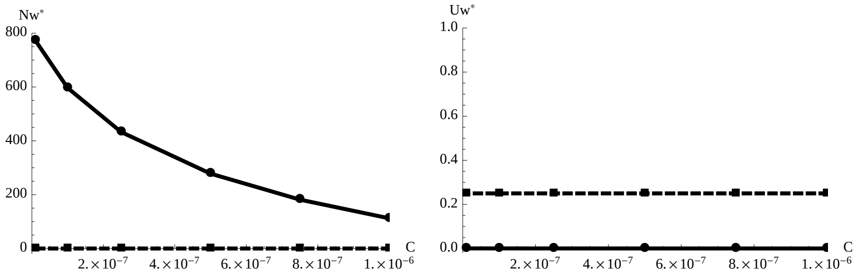
As with many real life security problems, the defender in this model is disadvantaged in that he takes only reactive measures against the phisher. Note that also we have omitted the case $c$ of Theorem 1 (i.e., when $\frac{1}{n_d-1} > \frac{R_w}{R_s} > \frac{1}{n_d}$), a relatively small region with points of discontinuity, for simplicity.

## 4 Analysis

We analyze using the CBP game three different scenarios: (i) the hypothetical case of perfect detection of phishing attacks, i.e., $P_d = 1$, (ii) $P_d < 1$ and is exogenously determined, and (ii) $P_d < 1$ and is endogenously influenced by the number of phishes the attacker creates.

**Perfect Phish Detection.** Let us start with the hypothetical case where the probability of detection, $P_d = 1$. Figure 2 plots the optimal number of additional phishing attacks $n_w^*$ that the phisher will launch depending on cost $c$, knowing that all newly created phishes will be detected by the defender. Note that this is exactly the scenario analyzed in [10], and that the dashed and solid lines plot the case $a$ and $b$ of Theorem 2 respectively. When the resource asymmetry is small (with $\frac{2}{n_0} \leq \frac{R_w}{R_s} = \frac{1}{2}$, dashed line), the phisher optimally chooses *not* to create additional phishes. There is no advantage to further stretch the defender as the attacker, given his resources, is expected to win in equilibrium a proportion of battlefields equals $\frac{R_w}{2R_s} = \frac{1}{4}$ as shown in Figure 2(b).

However, when the resource asymmetry is large (with $\frac{2}{n_0} > \frac{R_w}{R_s} = \frac{1}{900}$, solid line), the phisher will create additional phishing attacks to reduce the ability of the defender in locking down all of them. Especially when cost $c$ (measured in terms of the normalized utility) is negligible, $n_w^*$ approaches $\frac{2R_s}{R_w} - n_0 = 800$ given $\frac{R_w}{R_s} = \frac{1}{900}$ and $n_0 = 1000$. Even so, interestingly, the utility of the phisher is still less than $10^{-3}$. Meanwhile, as $c$ increases (see Figure 2(a)), the optimal number of new phishing attacks $n_w^*$ quickly approaches zero.



(a) $n_w^*$ against $c$

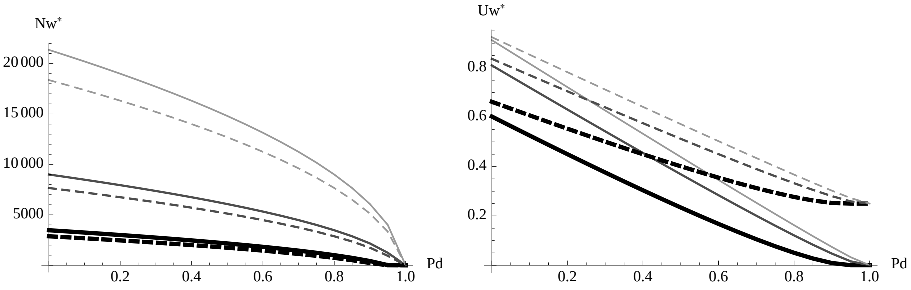
(b) $U_w^*$ against $c$

Fig. 2: The optimal new phishes $n_w^*$ and utility $U_w^*$ given $P_d = 1$. Solid and dashed lines plot the case where $\frac{R_w}{R_s} = \frac{1}{900}$ and $\frac{1}{2}$ respectively, with $n_0 = 1000$.

**Imperfect Phish Detection (Exogenous).** In practice, we can expect that a significant fraction of phishing attacks will get away undetected by the defender. The problem is exacerbated by non-sharing of data between different security vendors as observed in [15]. Figure 3(a) and 3(b) plot the optimal number of new phishes $n_w^*$ and the corresponding utility of the phisher $U_w^*$ depending on $P_d \in [0, 1]$. We assume that the phisher will be able to estimate $P_d$ based on past experience.

Let us first focus on the game between a resourceful (strong) phisher and the defender, with the resource asymmetry $\frac{2}{n_0} \leq \frac{R_w}{R_s} = \frac{1}{2}$ (as shown by the solid lines). Here, with $Pd < 1$, the phisher will now create additional phishes knowing that the defender will fail to detect some of the attacks, different from the case of perfect detection. The undetected phishes add on to the phisher's utility, which has a lower bound at $\frac{R_w}{2R_s} = \frac{1}{4}$. As for the game between a less resourceful (weak) phisher and the defender given a large resource asymmetry of $\frac{2}{n_0} > \frac{R_w}{R_s} = \frac{1}{900}$ (as depicted by the dashed lines), observe that the optimal numbers of new phishing attacks are now much higher than 800, the upper bound for the case of perfect detection.

Another interesting observation is that the utility gap between a strong and a weak phisher reduces as $P_d$ decreases from 1 to 0. Improving on $P_d$ thus will hurt a weak phisher, but has less impact on a strong phisher as he can leverage on his



(a) $n_w^*$ against $P_d$

(b) $U_w^*$ against $P_d$

Fig. 3: Optimal number of new phishes to create $n_w^*$ and the corresponding optimal utility $U_w^*$. Solid and dashed lines plot the case where $\frac{R_w}{R_s} = \frac{1}{900}$ and $\frac{1}{2}$ respectively, with $n_0 = 1000$. The effect of a decreasing cost $c$ going from $5 \times 10^{-5}$ to $1 \times 10^{-5}$ and $2 \times 10^{-6}$, measured in terms of the normalized utility, is depicted by the thick-black, normal-black and thin-gray lines, respectively.

resources (technologies, infrastructure, manpower, etc.) to resist the takedown of some of his phishes. The trend also suggests that an attacker will optimally vary his strategies to create more phishes when $P_d$ is low, but strive to increase his resources as $P_d$ increases.

Regardless of the extent of resource asymmetry, an increased cost (see the thick-dark lines versus the thin-gray lines) reduces both the optimal number of phishes and the utility of the phisher. But, somewhat counter-intuitively, the lower the detection probability, the more phishes the attacker will want to create. An attacker does not settle with having a fraction of undetected phishes, but will exploit the weakness of the defender in detecting all phishes and create even more phishes to increase his utility.

Another counter-intuitive and interesting finding is that in fact it is optimal for a less resourceful phisher to create more new phishes (than if he is a resourceful phisher) in equilibrium. This can be seen in Figure 3 where the solid lines ($\frac{R_w}{R_s} = \frac{1}{900}$) remain above the dashed lines ($\frac{R_w}{R_s} = \frac{1}{2}$) for all different costs $c$. This is surprising as large-scale phishing attacks are more often associated with resourceful attackers such as the 'Rock Phish' and 'Avalanche' gangs empirically.[15] There could be several reasons to this. First, while the 'Avalanche' phishes can be recognized easily with their distinctive characteristics, we do not know if the bulk of other phishing attacks are not related (carried out by a single organization) for sure. Secondly, could there be really a 'tragedy of the commons' due to the a large number of phishers (as described in [9]) that has forced the less resourceful attackers out of the phishing endeavor? We note that analyzing the effect of competition between several phishers would be an interesting extension

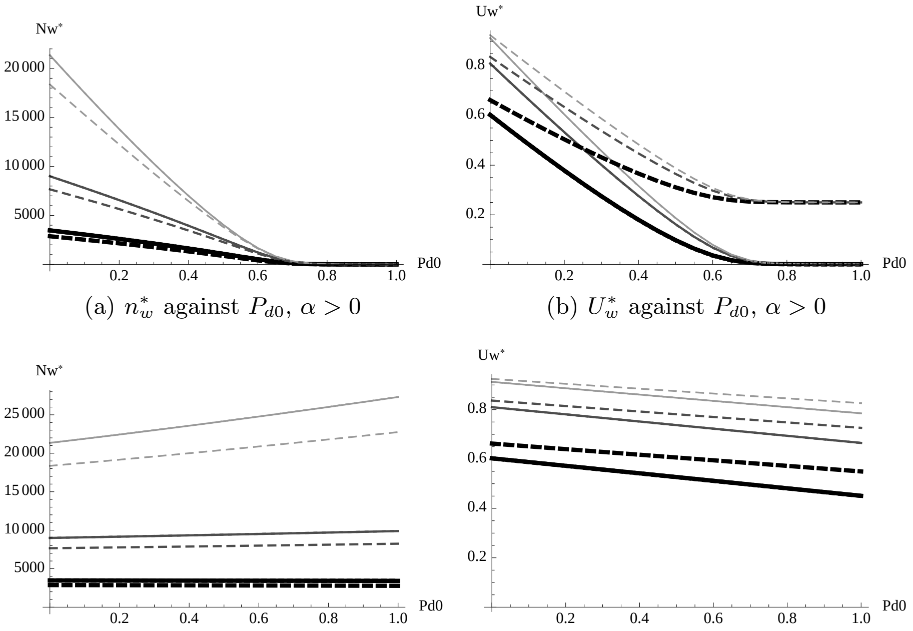
[15] For example, the 'Avalanche' gang was responsible for 84,250 out of 126,697 (66%) phishing attacks recorded by the APWG in the second half of 2009.

to our current model. Another more likely explanation would be that most of the phishing attacks are in fact detectable by the defender today, forcing the less resourceful attacker to gain too little utility to be profitable (observe that $U_w^*$ for the less resourceful attacker is almost zero as $P_d \to 1$ in Figure 3(b)). Furthermore, having a large number of phishes can also increase the probability of detection by the defender. We analyze the case when $P_d$ depends on the $n_w$ in the next section.

**Imperfect Phish Detection (Endogenous).** Let us model the effective $P_d$ to depend on the number of phishes an attacker creates with a simple formulation:

$$P_d = P_{d0} \times (n_w)^\alpha$$

where with $\alpha = 0$, we thus have the exogenous case as discussed in the previous section. The interesting analysis here is when $\alpha \neq 0$ as depicted in Figure 4.



(a) $n_w^*$ against $P_{d0}$, $\alpha > 0$

(b) $U_w^*$ against $P_{d0}$, $\alpha > 0$

(c) $n_w^*$ against $P_{d0}$, $\alpha < 0$

(d) $U_w^*$ against $P_{d0}$, $\alpha < 0$

Fig. 4: Optimal $n_w^*$ and $U_w^*$ when the effective probability of detection, $P_d = P_{d0} \times (n_w)^\alpha$. Graphs $a$ and $b$ plot the case where $\alpha = 0.05 > 0$, while graphs $c$ and $d$ plot the case of $\alpha = -0.2 < 0$. Solid and dashed lines plot the case where $\frac{R_w}{R_s} = \frac{1}{900}$ and $\frac{1}{2}$ respectively, with $n_0 = 1000$. The effect of a decreasing cost $c$ going from $5 \times 10^{-5}$ to $1 \times 10^{-5}$ and $2 \times 10^{-6}$ is depicted by the thick-black, normal-black and thin-gray lines, respectively.

There are many examples where increasing the number of phishing attacks (battlefields) can lead to a higher detection rate by the defender. For instance, the way the 'Rock Phish' and 'Avalanche' gangs hosted a number of phishing attacks (i.e., different phishing pages targeting different brands) using the same domain name[16], while reducing cost, increases the chance that all these phishes (battlefields) will be detected and taken down altogether. An attacker who register multiple domains for phishing purposes may also risk leaving visible patterns in the WHOIS database that is being used by the defender to identify and suspend suspicious domains quickly.[17]

As shown in Figure 4(a) and 4(b), both the $n_w^*$ and $U_w^*$ curves are now steeper than before. The optimal number of additional phishing attacks to create quickly approaches zero as $P_{d0}$ increases. Other than that, the main results from the case of exogenous detection probability (where $\alpha = 0$) remain applicable. First, it is optimal for a weak phisher to create more phishes than a resourceful attacker. The lower the detection probability is the more phishes will an attacker create. Also, improving the baseline detection technologies ($P_{d0}$) hurts a weaker phisher more than a stronger phisher.

It is harder to think of some practical examples where an increased number of phishes helps to reduce the effective detection rate by the defender (i.e., with $\alpha < 0$). A possible but *unlikely* scenario would be if the phishing attacks that a phisher creates cannot be correlated to each other, and that the larger number of attacks stretch the defender's capability in detecting all of them. We include the plots of optimal $n_w^*$ and $U_w^*$ under such scenario in Figure 4(c) and 4(d) for reference purposes. Notice that the optimal utility of the phisher is now bounded only by the cost of creating new phishes.

## 5 Discussion: Implications to Anti-Phishing Strategies

The success of anti-phishing defense depends on a number of interacting variables. As captured in our model, increasing the cost of creating new phishes $c$, improving the detection rate of new phishes $P_d$, as well as, increasing the resource asymmetry between the defender and phisher, $\frac{R_s}{R_w}$ are all crucial factors to be considered.

Increasing the cost for creating new phishes will hurt the attacker especially a weak phisher, who has no resources to resist the prompt removal of his phishes. Raising the cost (both in financial and procedural terms) for registering a domain name can therefore help, but only to a certain extent. Take the decision by CNNIC to make the registration of domain names more restrictive for example, the number of `.cn` phishing domains dropped, but phishing attacks on Chinese institutions remained high as phishers shifted to use other domain names such as `.tk` and the co.cc subdomain service (see [3] page 5). Phishers would also usually

[16] A typical 'Avalanche' domain often hosted around 40 phishing attacks at a time [2].

[17] APWG reported that attackers often utilize a single or small set of unique names, addresses, phone numbers, or contact email addresses to control their portfolio of fraudulent domain names [1].

register new domains using stolen credit cards. Furthermore, studies have found that a larger percentage of phishing attacks (80%) are actually performed using compromised web servers of innocent domain registrants (see e.g., [2,3,17]). To raise the cost $c$ will thus involve patching a large number of vulnerable servers, which is challenging if not impossible without a proper incentive plan.

A more effective alternative is hence to focus on improving the detection rate of new phishes. While automated spam filters help to detect potential phishing URLs, the 'Rock Phish' gang, for example, used GIF image in phishing email to evade detection. The popularity of URL shortening services and wall postings on online social networks add up to the challenge of detecting all phishing advertisements. Calls to share the phishing data in the anti-phishing industry have been made before (e.g., in [15]), but sharing can also create concerns as takedown companies leverage on their phishing data for competitive edges. Here, we see a room to employ and better coordinate the crowds to help improving the detection probability. Collecting user reports against potential phishes (or potentially harmful sites), without necessarily demanding from them higher skilled tasks such as evaluating if a phish is valid (or that a site is secure), can already be helpful.

Naturally, the value of data sharing and crowd-based phish-reporting will depend on the state of information asymmetry (i.e., the detection probability $P_d$). As can be seen in Figure 3, an 'intelligent' phisher will leverage on a large number of phishes for optimal utility when $P_d$ is low. Meanwhile, as $P_d \rightarrow 1$, a phisher will improve his utility by increasing his resources to match the defender's. This includes, for example, to gain access to a botnet infrastructure so to prolong the uptime of his phishes. Should a good estimate of $P_d$ is available, the defender can thus decide whether to prioritize on increasing the cost of creating new attacks (to reduce the number of phishes the attacker can create), or to prioritize on disrupting the channels a phisher can increase his resources (e.g., access to a botnet infrastructure, malicious tools, the underground market to monetize stolen credentials, or domain resellers with shady practices), accordingly.

## 6 Conclusions

We have proposed the Colonel Blotto Phishing (CBP) game to help better understanding the dynamics of the two-step detect-and-takedown defense against phishing attacks. We gained several interesting insights, including the counter-intuitive result that it is optimal for the less resourceful attacker to create even more phishing attacks than the resourceful counterpart in equilibrium, and that the attacker will optimally vary his strategies to either increase the number of phishes or to focus on raising his resources depending on the detection probability. We then discussed the implications to the anti-phishing industry.

Capturing the conflicts between an attacker and a defender with asymmetric resources and information, it is our hope that the CBP game can be eventually used to analyze other interesting problems, including measuring the effects of competition between multiple phishers, and the benefits of cooperation between multiple takedown companies. We also see the suitability of the CBP game to be applied to web security problems in general. Indeed, various web security problems, including malicious sites, illegal pharmacies, mule-recruitment and so fourth, are currently mitigated through a detect-and-takedown process similar to in the anti-phishing industry.

**Future Work.** Like other stylized models, the CBP game can be extended in several directions. A potential extension is to include the time dimension into the game, for example, using repeated games to model the uptime of a phish, which is often used to measure the damage caused by phishing activities. Using the variants of the classic Colonel Blotto game, such as the non-constant sum version [19] in which players might optimally choose not to expend all their resources, may also yield interesting results. We note that it may be interesting also to test our CBP model through experimental studies. Existing studies as conducted in [4,5,7,12] have largely found that subjects were able to play the equilibrium strategies of the classic Colonel Blotto game, with the weak and strong players adopting the 'guerrilla warfare' and 'stochastic complete coverage' strategies respectively. Testing how the subjects will play our two-stage CBP game can be an interesting future work.

## Acknowledgment

This research was supported in part by the National Science Foundation under award CCF-0424422 (TRUST).

## References

1. Anti-Phishing Working Group (APWG). Advisory on utilization of whois data for phishing site take down. `http://www.antiphishing.org/reports/apwg-ipc_Advisory_WhoisDataForPhishingSiteTakeDown200803.pdf`.
2. APWG. Global phishing survey: Trends and domain name use in 2H2009. `http://www.antiphishing.org/reports/APWG_GlobalPhishingSurvey_2H2009.pdf`.
3. APWG. Global phishing survey: Trends and domain name use in 2H2010. `http://www.antiphishing.org/reports/APWG_GlobalPhishingSurvey_2H2010.pdf`.
4. A. Arad and A. Rubinstein. Colonel blotto's top secret files. Levine's Working Paper Archive 814577000000000432, David K. Levine, Jan. 2010.
5. J. Avrahami and Y. Kareev. Do the Weak Stand a Chance? Distribution of Resources in a Competitive Environment. *Cognitive Science*, pages 940–950, 2009.
6. E. Borel. La théorie du jeu les équations intégrales à noyau symétrique. *Comptes Rendus de l'Académie des Sciences*, 173:1304–1308, 1921. English translation by Savage, L. (1953) The theory of play and integral equations with skew symmetric kernels, Econometrica 21:97–100.
7. S. M. Chowdhury, D. K. J., and R. M. Sheremeta. An experimental investigation of colonel blotto games. CESifo Working Paper Series 2688, CESifo Group Munich, 2009.
8. O. A. Gross and R. A. Wagner. A continuous colonel blotto game. *RAND Corporation RM–408*, 1950.

9. C. Herley and D. Florêncio. A profitless endeavor: phishing as tragedy of the commons. In *Proceedings of the Workshop on New Security Paradigms (NSPW)*, pages 59–70. ACM, 2008.
10. D. Kovenock, M. J. Mauboussin, and B. Roberson. Asymmetric conflicts with endogenous dimensionality. Purdue University Economics Working Papers 1259, Purdue University, Department of Economics, Dec. 2010.
11. D. Kovenock and B. Roberson. Conflicts with multiple battlefields. Purdue University Economics Working Papers 1246, Purdue University, Department of Economics, Aug. 2010.
12. D. Kovenock, B. Roberson, and R. M. Sheremeta. The attack and defense of weakest-link networks. Working Papers 10-14, Chapman University, Economic Science Institute, Sept. 2010.
13. D. K. McGrath and M. Gupta. Behind phishing: an examination of phisher modi operandi. In *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, pages 4:1–4:8. USENIX Association, 2008.
14. T. Moore and R. Clayton. Examining the impact of website take-down on phishing. In *Proceedings of the 2nd APWG eCrime Researchers Summit*, pages 1–13, 2007.
15. T. Moore and R. Clayton. The consequence of noncooperation in the fight against phishing. In *Proceedings of the 3rd APWG eCrime Researchers Summit*, pages 1–14, 2008.
16. T. Moore and R. Clayton. The impact of incentives on notice and takedown. In M. Johnson, editor, *Managing Information Risk and the Economics of Security*, 2008.
17. T. Moore and R. Clayton. Evil searching: Compromise and recompromise of internet hosts for phishing. In *Financial Cryptography and Data Security*, volume 5628 of *LNCS*, pages 256–272. Springer, 2009.
18. B. Roberson. The colonel blotto game. *Economic Theory*, 29(1):1–24, Sept. 2006.
19. B. Roberson and D. Kvasov. The non-constant-sum colonel blotto game. CESifo Working Paper Series 2378, CESifo Group Munich, 2008.